NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

BC2406 Analytics I Final Report

HDB resale price predictions

Instructor: CoB (NBS) Liu Peng

Submitted by: Seminar Group 7 Team 5

Choi Seunghwan (U2021319B)

Liu Jia Wei (U1910228G)

Nicole Tan Si Jie (U2010799L)

Timmothy Yonathan (U2040650E)

Wilson Tan Junlong (U1822241C)

Table of Contents

1. Executive Summary

This report aims to provide an in depth analysis on the factors that affect HDB resale pricing with the use of different machine learning models. The analysis aims to help young couples plan their future housing arrangements by providing accurate and reliable predictions of resale prices.

The project is approached in 5 stages:

1. Data Preparation

Data Preparation involves the cleaning and exploration of the data. Data Exploration highlights the different characteristics and underlying trends within the data which will aid in the development and optimisation of the model. Data cleaning aims to detect and correct any inaccurate or missing values in the records from the dataset and format them appropriately such that it can be used to train the models properly.

2. Initial Modelling

Initial Modelling shows how the models will turn out without any optimisation of the models or any transformation of the data. It allows for the planning of the project going forward by surfacing any issues faced during the modelling process or any inadequacies of the initial model.

3. Additional data collection

Additional data collection gathers additional data from which a more robust model could be developed. These additional data are scraped from public data sources, and feature creation is done to generate additional columns to feed into the new model.

4. Subsequent Modelling

The subsequent modelling of the data aims to optimise each model type with the new data collected.

5. Analysis

The results of the 3 models are compared to one another to determine the best model for the prediction of HDB resale flat prices. Further analysis on the business problem and limitations are provided.

2. Business understanding

2.1. Background

In Singapore, there are mainly 3 different housing options: public, hybrid and private properties. Over 80% of Singaporeans live in public housing, which is managed by the Housing Development Board (HDB) and are colloquially known as HDBs (H. D. B., 2021). Within public housing, couples usually have one of two choices, Build To Order flats (BTO) or Resale flats.

"Build to order (BTO) flats are HDB flats where construction will commence only if 65-70% of the apartments in the flat have been booked. The construction will be aborted if this requirement is not met." (PropertyGuru, 2017, para.1) Couples must be at least 21 years old when applying for BTO, and the whole process of applying for BTO from the launch to key collection will span a long period of about 5-6 years

Resale flats are HDB flats that are currently owned by someone else, who have already lived in for at least the minimum occupation period. The minimum occupation period is the period of time (usually five years) that you're physically required to live in your flat before you can sell or rent it out, or buy another private property. Couples buying resale flats are able to collect their keys once the transaction is completed.

2.2 Business Problem

University students are currently at the stage of life where many are planning for their future housing arrangements. However, when considering housing options, there are multiple factors that a couple has to take into consideration such as pricing, maximisation of grants, lead time and credit score. BTO flats are extremely popular among university students due to their low prices, greater chance for capital appreciation and accessibility to higher housing grants.

These factors, coupled with the uncertainty of HDB resale prices in the future lead to couples often considering resale HDB flats as a last resort (CNA, 2021). Therefore, it is to no surprise that BTO flats are oversubscribed, especially for locations that are highly popular (The Straits Times, 2021) where BTO applications can shadow available flats by over 20 times. Consequently, many young couples are often rejected for their first or even second application which, coupled with the long lead time for BTOs (5-6years) can lead to tremendous stress and anxiety.

As such, the team aims to provide young couples with more information regarding the price and characteristics of resale flats which they can rely on to make an informed decision on their future housing arrangements and work out their finances.

3. Initial Data Preparation

3.1 Description of Data's Characteristics

The data was downloaded from kaggle and contains data regarding the sale of Housing Development Board (HDB) resale flats from 1990 to 2020. The data was split into 5 csv files with a combined total of 826,581 transactions. There are 10 columns in the data as seen below:

1. month - month and year of transaction
2. town - town in which the flat resides
3. flat_type - an indication of the size and number of room in a flat
4. block - block number of the flat
5. street name - street name of the flat
6. storey_range - indication of how high or low the flat is in the block
7. floor_area_sqm - floor area of flat in square meters
8. flat_model - model of flat
9. lease_commence _date - year the lease started for the flat
10. resale_price - nominal resale price of the flat

Elaboration about the different types of flat models can be found in *Appendix A*.

3.2. Initial Data Cleaning

3.2.1 Time period of data

The time period used for the project is between the years 2012 and 2020 which are deemed to be more relevant. The relevant files were combined with the following columns: 'town', 'flat_type', 'storey_range' and 'flat_model' factored as they have a fixed and known set of possible values and should be treated as categorical for statistical modelling. The various flat models were also generalized to by grouping flat models with a high degree of similarity for a more insightful comparison.

3.2.2 Missing Values

There are missing values in the column 'remaining_lease' which was computed with the following formula:

remaining lease = lease commence date + 100 - (year + month/12)

The 'month' column was split to obtain the variables year and month for the computation.

3.3.3 CPI Adjusted Resale Price

Given that the data was retrieved from different years, it has not accounted for inflation which might affect the precision of the models. Therefore, the consumer price index (CPI), which is the most widely used measure of inflation, was merged with the flat resale price data to compute the resale price adjusted for inflation. This ensures that any differences in the resale prices are not attributable to inflation which is not a variable that is being considered.

The cleaned data was stored in 'CLEANED_COLLECTED_hdb_data.csv', which will be used for any subsequent analysis.

3.3 Initial Data Exploration

First, the distribution of resale flats in the different towns in Singapore was observed via a bar graph as seen in Figure 1 *(Appendix B)*. From the figure, the largest number of resale flats are populated in towns such as Jurong West, Woodlands and Hougang while the lowest number of resale flats are in the Central Area, Marine Parade and Bukit Timah respectively. Generally, the former towns are considered to be non-mature estates, which tend to be newer and further away from the city centre. Conversely, the latter tend to be mature estates and are significantly more costly. Afterwards, the resale prices was compared across the different towns as shown in Figure 2 *(Appendix B)* using a violin plot. As seen from Figure 2 (*Appendix B),* there is a general increase in the median resale prices, with towns such as Jurong West, Woodlands and Hougang being on the lower end, and towns such as the Central Area, Marine Parade and Bukit Timah being on the higher end of the spectrum. This trend runs parallel to Figure 1 *(Appendix B)*. As such, we can conclude that generally, towns with higher resale prices have a lower population of resale flats and vice-versa.

Next, the distribution of flat types were presented in a bar graph in Figure 3 *(Appendix B)* while corresponding resale prices were compared against different flat types using a jittered dot plot (Figure 4, *Appendix B)*.
Figure 3 shows that the 4-room flat type constitutes the large majority of resale flats in Singapore at 40% but despite the 4-room flat being the most common, they are neither the cheapest nor most expensive, showing a lack of correlation between distribution of flats and resale price.
From figure 4, it can be generally deduced that greater the flat type, the greater the resale price, with the 1-room flat type being the lowest resale price and the 5-room flat type being the highest.

Similar analyses were performed to compare resale price against storey range (Figure 5, *Appendix B*) and floor area per square metre (Figure 6, *Appendix B*) using box plot and scatter plot respectively.
The results were that higher storey ranges and greater floor area corresponds to higher resale flat prices. However, it was observed that a small number of 3-room flat types had a significantly greater floor area compared to even executive flats or multi-generation flats. Upon further investigation, it was found that these flats were of the flat model 'Terrace' (*Appendix A*) which are public landed properties, explaining the large floor area despite having few rooms.

Analysis to compare the resale prices across the flat models as seen in Figure 7 *(Appendix B)* was also performed using a box plot. From the figure, it can be seen that resale prices are quite volatile with regards to flat models. Therefore, if based solely on Figure 7, it is difficult to conclude if flat models have a significant influence over resale prices.

4. Initial Modelling

The approach for modelling is based on 3 machine learning models of varying complexities: Linear Regression, Classification And Regression Trees (CART) and Random Forest.

For the training of the model, variables relating to transaction date such as 'year' and 'month' were removed given that the resale price has already been adjusted for inflation and they are unlikely to affect the resale price. Variables relating to address such as 'address', 'block', 'street name', 'latitude' and 'longitude' were also removed due to the sheer amount of factors involved. The variable 'town' would be used to account for any geographical influences instead.

4.1 Linear Regression

Linear regression is one of the most basic and commonly used predictive modelling techniques. The aim of linear regression is to find a mathematical equation for a continuous response variable Y as a function of one or more independent variables, capturing the magnitude of effect that each independent variable has on Y. Given the multitude of variables, Multivariate Linear Regression was conducted to predict the prices of resale flats.

4.1.1 Methodology

The dataset was split into a train set and a test set with a 70:30 split. The train-test split is commonly used to train and evaluate the model separately to obtain an unbiased evaluation of the model, reducing the chance of overfitting. The regression model was run using the lm() function in R with the resale price used as the Y(dependent) variable.

4.1.2 Evaluation

The metrics used in evaluating the performance of the model were Adjusted R-squared, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) and are shown in Figure 8 below. The team feels that the model has a decent performance but can be improved further.

```
Residual standard error: 53430 on 118795 degrees of freedom
Multiple R-squared:  0.8574,    Adjusted R-squared:  0.8573
F-statistic:  9397 on 76 and 118795 DF,  p-value: < 2.2e-16

> # RSME: 52,868.8
> sqrt( sum( (actual_pred_trial$actuals - actual_pred_trial$predicted)^2 ,
al) )
[1] 52868.8
> # Mean Absolute Error(MAE) : 40,385.24
> mae(actual_pred_trial$actuals, actual_pred_trial$predicted)
[1] 40385.24
```

Figure 8

4.2 Decision Tree Regression (CART for Continuous Y)

Decision Tree Regression builds a tree model in which the data is sequentially broken down into smaller subsets following association rules. For continuous Y, the metric used to determine the best split is based on the sum of squared errors and the model predicts the Y value based on the mean for each split.

4.2.1 Methodology

The train-test split of 70:30 was also done for CART modelling with K-fold cross validation being applied on the test set. The features used for the modelling were selected based on domain knowledge and the variables chosen were town, flat type, storey range, floor area, flat model and remaining lease for the resale flats.

4.2.2 Evaluation

	Cp Value	RMSE	MAE	MAPE
Optimal Tree	4.443322e-06	38785.05	28006.82	0.0638134
Pruned Tree 1	1.000000e-03	56766.56	42570.35	0.09817055
Pruned Tree 2	5.000000e-03	67261.81	50683.34	0.117704

Table 1

The metrics used for evaluation of the CART model are RMSE and MAE which are also used for evaluation of linear regression together with another metric Mean Absolute Percentage Error (MAPE). The values are summarized in Table 1 above.

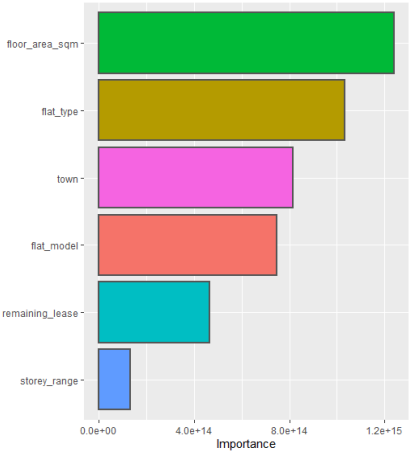


Figure 9

The variable importance plot for the optimal tree is shown in Figure 9, with floor area being the most important, followed by flat type. The variable importance for the pruned trees follow the same ranking as the optimal tree.

4.3 Random Forest Regression

Random forest regression is a supervised learning algorithm that uses an ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction as compared to a single model. A single decision tree has high variance as it tends to overfit to the data.

However, through bagging and ensembling, random forest is able to reduce the variance of each tree by combining them. A java-based 'h2o' library was used for faster modeling.

To build the most optimal model, the 'h2o.grid' function was used to train multiple models with different parameters, taking advantage of the 'h2o' library's ability which allows for different optimal search paths in the grid search.

- 'mtries' specifies the number of columns to randomly select at each level. A vector of positive integers lower than the n_features was assigned.
- 'min_rows' specifies the minimum number of observations for a leaf in order to split. A value too low will cause underfitting and a value too big will cause overfitting. Hence, a vector of values 10, 12, 15, 18, 20 were used.
- 'nfold' = 10 was used for k-fold cross validation, the same validation method as explained in the previous section (Linear Regression).
- 'Search_criteria' has a default method of "cartesian" search which covers the entire space of hyperparameters combinations but it would take extremely long to complete the search. Therefore, we used "RandomDiscrete", which will jump from one random combination to another once a certain level of improvement has been made or a certain amount of time has been exceeded. The stopping criteria for model selections were as following:
    - 'stopping _metric' = "rmse"
    - 'stopping_tolerance' = 0.01
    - 'stopping_rounds' = 10

This means that the grid search will stop if none of the last 10 models had a 1% improvement in root mean-squared error.

- Within each model, to implement early stopping,
    - 'stopping_metric' = "mse"
    - 'stopping_tolerance' = 0.005
    - 'stopping_rounds' = 10

The best model, based on the R-squared value, was chosen using the 'h2o.getModel()' function. Metrics reported on the training data were as shown in Figure 10 below, with a R-squared value of 0.9212679. For the prediction on the test set, the result is shown in Figure 11, with R-squared value of 0.9233841.

```
H2ORegressionMetrics: drf
** Reported on training data. **
** Metrics reported on Out-Of-Bag training samples **

MSE:  1581993283
RMSE:  39774.28
MAE:  29611.11
RMSLE:  0.08639187
Mean Residual Deviance :  1581993283
```

Figure 10

```
H2ORegressionMetrics: drf

MSE:  1553896437
RMSE:  39419.49
MAE:  29295.87
RMSLE:  0.08563381
Mean Residual Deviance :  1553896437
```

Figure 11

5. Subsequent Data Collection

5.1 Need for additional data

Based on the initial modelling result, the team feels that the modelling can be improved as the variables included do not cover all the factors that would significantly affect the resale price of HDB flats. In order to build a model with a more holistic and accurate prediction of the resale price, variables such as the distance to amenities and the number of amenities within a specified distance need to be collected. The variables will be added to the data and this whole process is known as feature creation.

5.2 Methodology

The first step would be to obtain the list of amenities such as schools, public transport, malls and hospitals in Singapore and their respective addresses. Some of the data was scraped from websites such as Wikipedia and moe.gov.sg while the rest were obtained from data.gov.sg or called using data.gov.sg's API. The addresses were then used to call an API from onemap.sg to obtain the full addresses and coordinates of the amenities. The API was also used to obtain the coordinates of the HDB resale flats. The coordinates of the HDB flats were compared against the coordinates of the amenities to compute the distance to amenities as well as the number of amenities nearby. Finally, the data was merged with the cleaned HDB data for further use.

5.3 Additional data exploration

Amenities-wise, a map plot for geographic visualisation was generated. In the process, google maps was selected as the resource for mapping. The relevant amenities were classified into healthcare (hospital and polyclinics), travel (LRT and MRT), food and retail (hawker centres, shops, markets) and schools and parks respectively. As illustrated in Figure 11 *(Appendix B),* there is little distribution of healthcare amenities across Singapore, with a large majority concentrated in the central regions. In contrast, there is a vast number of MRT and LRT stations spread all around Singapore as seen in Figure 12 *(Appendix B)*. However, it is noted that there are some areas such as Lim Chu Kang, Boon Lay and Changi which do not have any public transport amenities, so they may be considerably less convenient.

In addition, as seen from Figure 13 *(Appendix B)*, there is a fair distribution of food and retail options around the country. Specifically, the central-to-south region has a greater concentration of these options. This is due to the region largely consisting of more mature estates such as Bukit Merah and the Central Area, which is also prided as the country's shopping precinct. Similarly, in Figure 14 *(Appendix B)* and Figure 15 *(Appendix B)* respectively, there is easy access to many school and park facilities as they are easily seen to be dotted across the local landscape.

6. Subsequent Modelling & Evaluation

6.1 Linear Regression

Given the additional data about amenities, linear regression was run again with hopes that the new variables would help to build a more accurate model.

6.1.1 Data Preparation

The data has to be prepared before being used to train the model. Similar to the initial modelling, columns containing information relating to the address and date of transaction were removed together with any duplicate rows. Furthermore, Figure 5 (*Appendix B*) suggests that the frequency of records for storey_range '31-35' and '36-40' might be insufficient for them to be statistically significant. An iteration was run to remove all levels of storey_range with the number of records under 20.

The train-test split is maintained at 70:30 and the results of the model are shown in Figure 16. The adjusted R-squared has increased from 0.8573 in the initial model to 0.8843.

```
Nearest_station               -5.092e+04  5.000e+02 -101.833  < 2e-16 ***
Number_of_stations_in_2km       1.410e+03  7.404e+01   19.045  < 2e-16 ***
Nearest_supmarket               1.462e+03  4.107e+02    3.561  0.00037 ***
Number_of_supmarkets_in_2km     1.263e+03  8.543e+01   14.786  < 2e-16 ***
Nearest_park                  -5.201e+03  5.016e+02  -10.369  < 2e-16 ***
Number_of_parks_in_2km        -6.758e+01  5.100e+01   -1.325  0.18513
Nearest_polyclinic            -1.366e+03  3.198e+02   -4.270 1.95e-05 ***
Number_of_polyclinic_in_2km     2.169e+03  3.848e+02    5.636 1.74e-08 ***
Nearest_school                  1.656e+04  8.824e+02   18.772  < 2e-16 ***
Number_of_schools_in_2km        1.069e+03  6.442e+01   16.591  < 2e-16 ***
Nearest_mall                  -1.435e+04  5.081e+02  -28.241  < 2e-16 ***
Number_of_shopmalls_in_2km    -1.310e+03  7.992e+01  -16.396  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 48090 on 118428 degrees of freedom
Multiple R-squared:  0.8844,    Adjusted R-squared:  0.8843
F-statistic: 1.041e+04 on 87 and 118428 DF,  p-value: < 2.2e-16
```

Figure 16

6.1.2 Optimising Model

After the creation of the first model, we seeked to optimise the model for the best results as illustrated below in Figure 17 and 18 respectively. The metrics used in evaluating and optimising the performance of the model were Adjusted R-squared, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), similar to what was used for the initial models.

flat_modelAdjoined flat	7.903e+04	2.803e+04	2.819	0.00481	**
flat_modelApartment	3.392e+04	2.785e+04	1.218	0.22323	
flat_modelDBSS	1.243e+05	2.785e+04	4.465	8.03e-06	***
flat_modelExecutive Maisonette	1.403e+05	2.938e+04	4.777	1.78e-06	***
flat_modelImproved	-8.903e+03	2.781e+04	-0.320	0.74890	
flat_modelMaisonette	6.642e+04	2.785e+04	2.385	0.01708	*
flat_modelModel A	-8.817e+03	2.781e+04	-0.317	0.75121	
flat_modelModel A2	-1.288e+03	2.784e+04	-0.046	0.96310	
flat_modelMulti Generation	NA	NA	NA	NA	
flat_modelNew Generation	-1.950e+03	2.782e+04	-0.070	0.94412	
flat_modelPremium Apartment	2.887e+03	2.781e+04	0.104	0.91733	
flat_modelSimplified	-2.381e+03	2.782e+04	-0.086	0.93179	
flat_modelStandard	1.027e+04	2.782e+04	0.369	0.71201	
flat_modelTerrace	3.707e+05	2.839e+04	13.056	< 2e-16	***
flat_modelType S1S2	1.612e+05	2.808e+04	5.742	9.37e-09	***

Figure 17

flat_type2 ROOM	7.716e+03	6.286e+03	1.228	0.21964	
flat_type3 ROOM	4.283e+04	6.217e+03	6.890	5.61e-12	***
flat_type4 ROOM	6.641e+04	6.359e+03	10.444	< 2e-16	***
flat_type5 ROOM	8.506e+04	6.577e+03	12.934	< 2e-16	***
flat_typeEXECUTIVE	8.750e+04	6.830e+03	12.812	< 2e-16	***
flat_typeMULTI-GENERATION	2.276e+05	2.953e+04	7.706	1.31e-14	***

Figure 18

Based on the first model, the variable ‘flat model’ was shown to be statistically insignificant and was thus removed from subsequent models (Figure 17). A second model was run without ‘flat_model’ as a variable. The resulting model shows that the 2 room flat type was also statistically insignificant (Figure 18) and based on data exploration done previously, it was suspected that it might be due to insufficient data. A simple query was run (Figure 19) and it was seen that ‘1 room’ and ‘multi-generation’ flats make up less than 0.1% of the data. The data for 1 room and multi generation flats were thus removed as it was unlikely that given the few examples, they would be statistically significant. However, this would mean that the model would not be able to account for predictions where the flat types are 1 room or multi-generation. The stepAIC function was then run to choose a model using a stepwise algorithm but the resulting model does not show any improvement in metrics.

```
[1] "1 ROOM 90"
[1] "0.05 %"
[1] "2 ROOM 2145"
[1] "1.27 %"
[1] "3 ROOM 44483"
[1] "26.34 %"
[1] "4 ROOM 68280"
[1] "40.44 %"
[1] "5 ROOM 40365"
[1] "23.9 %"
[1] "EXECUTIVE 13433"
[1] "7.96 %"
[1] "MULTI-GENERATION 63"
[1] "0.04 %"
```

Figure 19

6.1.3 Multicollinearity

```
> vif(m3)
                                GVIF Df GVIF^(1/(2*Df))
town                   4.424309e+09 25        1.559254
flat_type              1.484850e+01  4        1.401072
storey_range           1.707705e+00 20        1.013469
floor_area_sqm         1.240955e+01  1        3.522719
lease_commence_date    6.120962e+02  1       24.740579
remaining_lease        5.907457e+02  1       24.305261
LATITUDE               5.963204e+01  1        7.722179
LONGTITUDE             1.033054e+02  1       10.163925
year                   3.190713e+01  1        5.648640
Nearest_hawker         1.597812e+01  1        3.997264
Number_of_hawker_in_2km 1.310543e+01  1        3.620143
Nearest_hospital       1.483313e+01  1        3.851380
```

Figure 20

Multicollinearity refers to the occurrence of high intercorrelations among two or more independent variables. Using the variance inflation factor (vif), we can estimate the extent of multicollinearity within the model. Multicollinearity can be an issue as it would be difficult to determine the individual effects of independent variables due to their high correlation with one another. In turn, this can reduce the accuracy of coefficient estimates and can cause it to be over sensitive to changes, resulting in a weaker model. Based on the values obtained from the vif function (Figure 20), we took out variables that have values higher than 5 such as lease commencement date, date, latitude and longitude. A fourth model was run and the metrics, where RMSE is 51,999.1 and MAE is 39,283.95, are shown below (Figure 21).

```
Residual standard error: 52170 on 118345 degrees of freedom
Multiple R-squared:  0.8635,    Adjusted R-squared:  0.8634
F-statistic: 1.118e+04 on 67 and 118345 DF,  p-value: < 2.2e-16

> sqrt( sum( (actual_pred$actuals - actual_pred$predicted)^2 , na.rm = TRUE ) / nrow(actual_pred) )
[1] 51999.1
> # RMSE: 51,999.1
> mae(actual_pred$actuals, actual_pred$predicted)
[1] 39283.95
> # MAE: 39,283.95
```

Figure 21

6.1.4 Label Encoding

Label Encoding was used on the categorical variables in an effort to improve model accuracy. However, the resulting models arising from label encoding were not ideal and had worse metrics. This can be attributed to the nature of the categorical variables. Label encoding is suitable for ordinal variables to capture any ordinal relationships within the model. However, the categorical variables for the dataset are nominal in nature and it would therefore be inappropriate to assign a nominal relationship to the variables. One-hot encoding would be more appropriate and is already used in linear regression by default.

6.1.5 Feature Scaling (for LR)

Feature scaling involves standardising or normalising the variables in the dataset. For regression models, scaling does not affect the statistical inference, but makes it easier to interpret the intercept term as the expected value of Y. As seen from Figures 22 and 23, the intercept value went from a negative number to a positive number.

```
> summary(m4)$coef
                          Estimate  Std. Error    t value
(Intercept)           -256096.1815 3583.36948 -71.467981
townBEDOK              -28915.7182 1403.96307 -20.595783
townBISHAN              73446.7993 1415.43490  51.889917
townBUKIT BATOK        -22648.3363 1977.29144 -11.454223
townBUKIT MERAH         45993.8047 2110.96049  21.788094
townBUKIT PANJANG      -71679.9936 1757.38328 -40.787911
townBUKIT TIMAH        178333.1612 3393.78739  52.546946
townCENTRAL AREA        88444.3974 2912.12332  30.371103
townCHOA CHU KANG      -90727.4651 1946.71882 -46.605326
townCLEMENTI            28914.6959 1618.50258  17.865091
townGEYLANG            -52919.4168 2215.22425 -23.888966
townHOUGANG             -4333.0751 1565.74316  -2.767424
townJURONG EAST         -8905.0067 2387.03175  -3.730577
townJURONG WEST       -108854.1044 1575.07876 -69.110261
townKALLANG/WHAMPOA     48784.2137 1858.01765  26.256055
townMARINE PARADE      100103.8944 2660.46463  37.626471
```

Figure 22

```
> summary(scaled_model)$coef
                          Estimate Std. Error    t value
(Intercept)           402917.4117 2120.2919 190.0292222
townBEDOK             -28204.3703 1406.2972 -20.0557679
townBISHAN             74694.2409 1413.8561  52.8301595
townBUKIT BATOK       -22054.8767 1980.1184 -11.1381606
townBUKIT MERAH        47841.1121 2106.6365  22.7097141
townBUKIT PANJANG     -70033.0506 1757.8032 -39.8412358
townBUKIT TIMAH       180099.7143 3338.3762  53.9482988
townCENTRAL AREA       86532.9494 2928.9802  29.5437127
townCHOA CHU KANG     -89828.4961 1944.6937 -46.1915902
townCLEMENTI           29701.7899 1628.2710  18.2413058
townGEYLANG           -52785.8162 2221.4050 -23.7623561
townHOUGANG            -2609.8135 1572.0485  -1.6601354
townJURONG EAST        -9745.2478 2384.8323  -4.0863451
townJURONG WEST      -109090.7327 1577.3086 -69.1625814
townKALLANG/WHAMPOA    50937.7969 1851.1341  27.5170752
townMARINE PARADE      99445.3722 2654.6728  37.4605835
```

Figure 23

6.1.6 K-Fold Cross Validation

Using the train-set approach where we hold out a part of the data for evaluation can introduce uncertainty to the model. This is due to the randomness of the split where the train set and the test set are unlikely to be fully representative of the data as a whole and thus run the risk of overfitting or underfitting. Using cross validation would minimise that risk by splitting the data into subsets to train and evaluate the model multiple times. The value of K chosen was 10 which has been shown empirically to yield test error rate estimates that suffer neither from excessively high bias nor from very high variance.

Comparing the metrics of the model from 10-fold cross validation shown in Figure 24 to that of Figure 21(without cross validation), it can be seen that the cross validation itself has minimal impact on model performance.

Figure 25 shows the residuals plot for the 10-fold cross validation and helps to visualise the performance of the model (specifically shows Adjusted R-squared). Given the strong correlation seen from the figure, it can be concluded that the model is accurate in predicting the resale price with the exceptions of a few outliers.

```
Residual standard error: 52320 on 118345 degrees of freedom
Multiple R-squared:  0.8628,    Adjusted R-squared:  0.8627
F-statistic: 1.11e+04 on 67 and 118345 DF,  p-value: < 2.2e-16

> #adjusted r-square = 0.8627
> sqrt( sum( (actual_pred_m5$actuals - actual_pred_m5$predicted)^2
[1] 51631.82
> # RSME: 51,631.82 (not ideal)
> mae(actual_pred_m5$actuals, actual_pred_m5$predicted)
[1] 39092.45
> # MAE: 39,092.45
```

Figure 24

Figure 25

6.2 Decision Tree Regression (CART for Continuous Y)

6.2.1 Data Preparation

The decision tree regression was conducted with the additional data feature that was mined to achieve a more accurate model. The additional features were:

- Nearest_hawker + Number_of_hawker_in_2km
- Nearest_hospital + Number_of_hospital_in_2km
- Nearest_station + Number_of_stations_in_2km
- Nearest_supmarket + Number_of_supmarkets_in_2km
- Nearest_park + Number_of_parks_in_2km
- Nearest_polyclinic + Number_of_polyclinic_in_2km
- Nearest_school + Number_of_schools_in_2km
- Nearest_mall + Number_of_shopmalls_in_2km

6.2.2 Optimising the Model

The same train-test split of 70:30 was used and a minimum split of 100 was specified in an attempt to control the tree such that it would not overfit. Minimum split refers to the minimum number of observations required before a split is carried out. Two techniques were performed to reduce the errors and improve the performance of the model.

6.2.3 Data Discretization

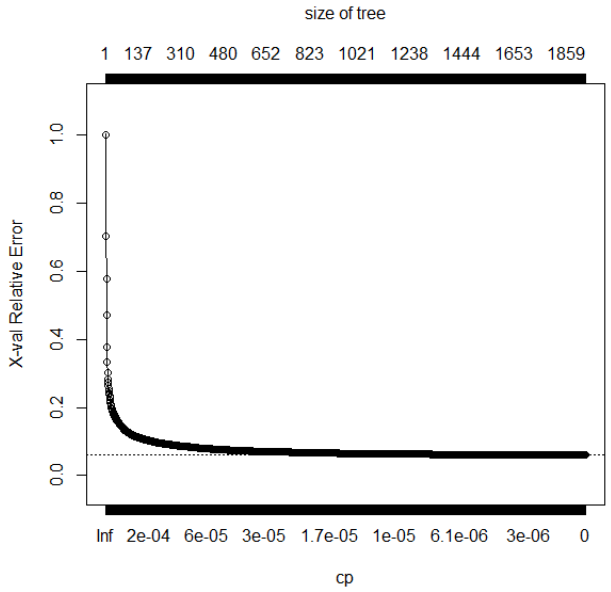
Discretization in theory is the act of reducing the range of values continuous variables take on by grouping themselves into bins. Discretization carried out on the floor area variable was found to slightly improve the model's performance. This variable was chosen as the resale price could be more sensitive to a certain range of the floor area, rather than the specific values. For instance, a floor area of 50 to 60 square metres could command the same level of influence over the resale price. Floor area below 50 square metres or above 60 square metres would, however, affect resale prices. This was achieved by splitting the floor area into 25 factors, where each range of values accounted for approximately 10 square metres. If the number of intervals were too high or too low, it would reduce the performance of the model. As such, a trial and error was performed which concluded that the best split for the range of floor area was into 25 parts.

6.2.4 Feature Scaling (for CART)

Normalization of the variable remaining lease was done by reducing the range of values to between 0 and 1. This is done by dividing the remaining lease column by 99 since the length of lease for HDB flats is known to be 99 years. Remaining lease cannot take on negative values and the minimum value is 0. It was found that normalizing the remaining lease feature slightly improved the model performance.

To evaluate the impact that the two techniques had on the performance of the CART model, the results of the optimal tree were compared between a model without discretization and

normalization and a modified model with both techniques performed. However, only the visualization for the modified model will be shown below.



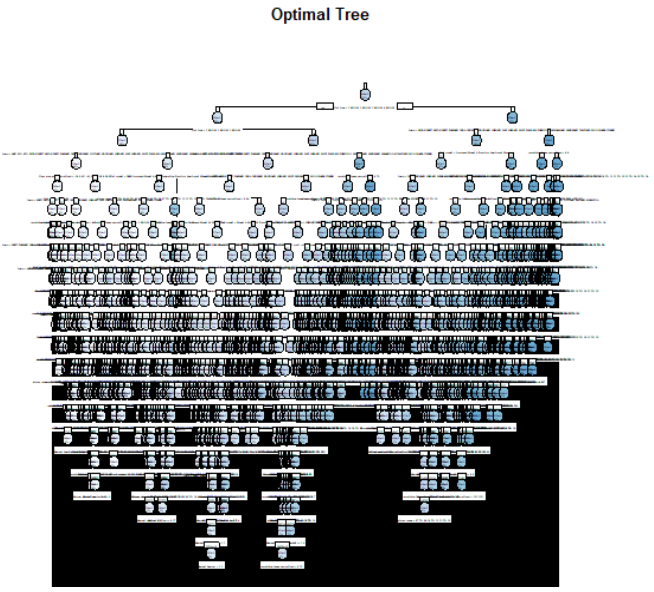
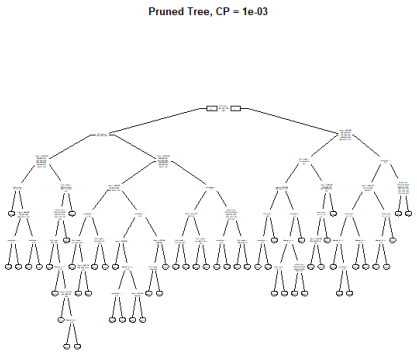


Figure 26

Figure 27

The maximal tree was trained and the complexity parameter (CP) was plotted, as seen in Figure 26 above. The optimal CP cannot be obtained by visual inspection and can only be obtained by referencing the cp_table. The optimal tree of the modified model plotted following the 1 standard error rule is shown in Figure 27 above. As the optimal tree is too complex, an arbitrary value of 0.001 and 0.005 was set as the new cp value to obtain a tree that is more generalized and user-friendly, as seen in Figures 28 and 29 below.



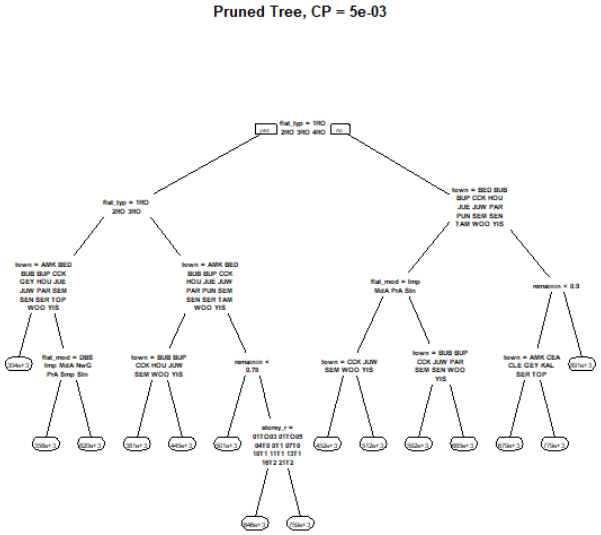


Figure 28

Figure 29

6.2.5 General Interpretation of terminal nodes

The decision tree starts from the root node and follows a set of decision rules until the terminal node is reached. Each internal node represents a test and each split represents the outcome of the test. The left side of the split represents a "Yes" to the decision rule at the root node while the right side represents a "No". A root node 1 would be split to node 2 if it fits the decision rule, otherwise it would be split to node 3. Similar splits for node 2 and 3 would be conducted until the terminal nodes are obtained. Therefore, the interpretation of a terminal node is the combined test condition of all the internal nodes leading to the terminal node. For instance, the combined test outcomes for node 1,2 and 4 would be required to predict the continuous variable Y at terminal node 8

6.2.6 Comparing the trees with varying CP values

	Min Split	Cp Value	RMSE	MAE	MAPE
Optimal Tree (Initial)	Default	4.443322e-06	38785.05	28006.82	0.0638134
Optimal Tree	Default	1.875444e-06	30702.24	22197.04	0.05090684
Optimal Tree	100	4.147897e-06	34878.15	25281.65	0.05778240
Modified Optimal Tree	100	4.298268e-06	34303.19	25058.32	0.05741735
Modified Pruned Tree 1	100	1.000000e-03	54818.42	41689.84	0.096490330
Modified Pruned Tree 2	100	5.000000e-03	68117.56	51755.25	0.12046560

Table 2

The metrics for evaluation of the models are summarized in Table 2 as illustrated. The optimal tree with a min split value of 100 results in a higher magnitude of error compared to the optimal tree with a default min split value. This reduces the chances of overfitting as the default min split could be too small when dealing with a large number of observations.

Between the optimal tree (min split =100) and the modified optimal tree in which selective feature discretization and normalization was performed, a slight improvement in metrics was observed. Amongst the three modified trees, the optimal tree has the best predicting power as it is the most complex. Comparing the modified optimal tree with the optimal tree (initial), the RMSE value was approximately lowered by 4,400 while the MAPE improved from 6.38% to 5.74%.

6.2.7 Plotting the graph of predicted vs actual resale price with varying cp values

The predicted price was plotted against resale price for graphical comparison as shown in Figure 30. The plots are arranged in ascending order, with the top plot having the lowest cp value of 4.15e-06 (truncated to 2 decimal places) and the bottom plot having the highest cp value of 5.00e-03. Points lying on the line y = x represent predictions of resale price that

agree with the actual resale price. Down the plots, the distance between the points from the reference line y = x deviates, signifying a larger margin of error.

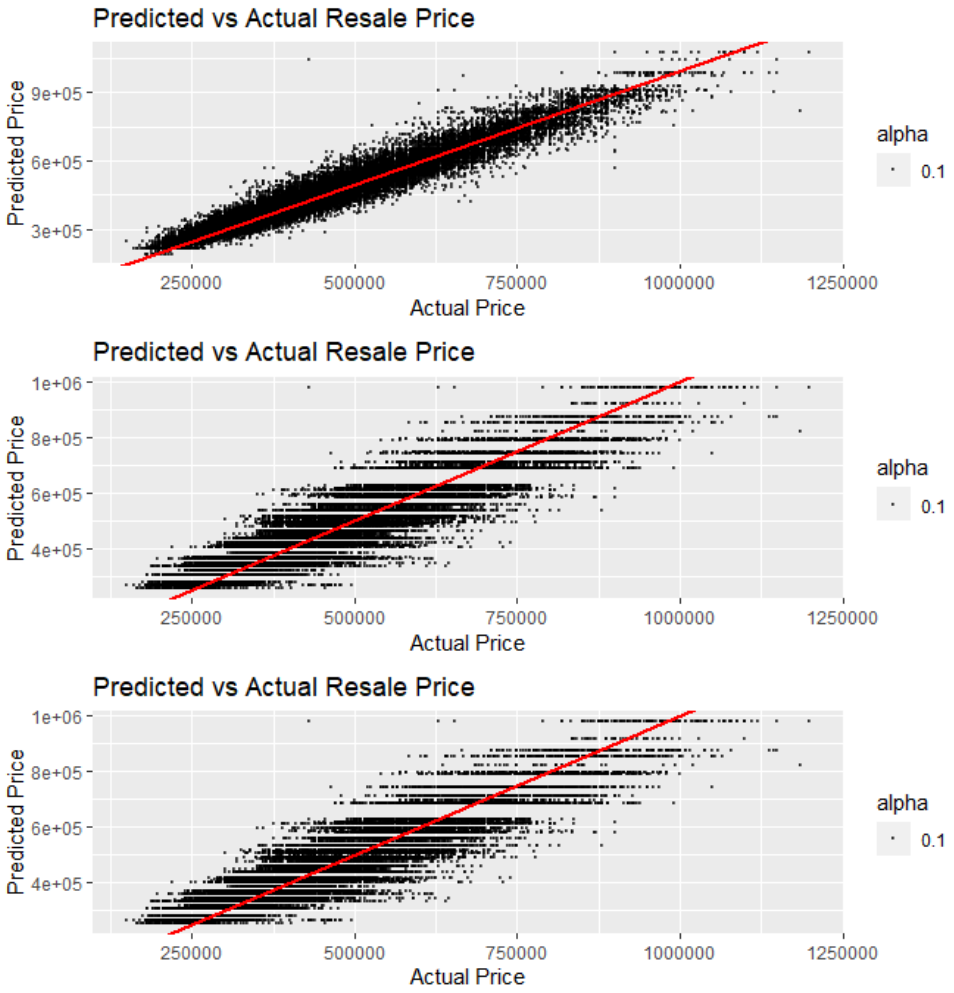


Figure 30

6.2.8 Variable Importance of each tree

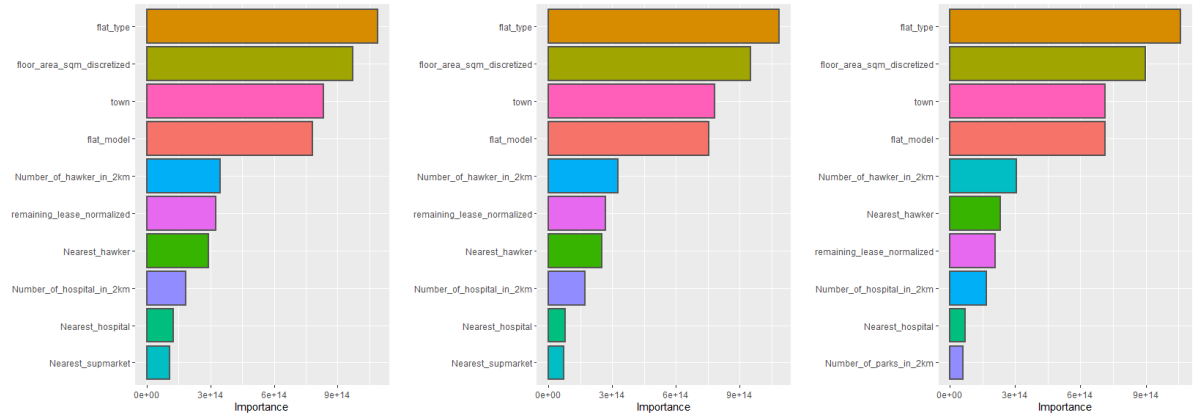


Figure 31

As seen in Figure 31, the variable importance of the optimal tree is plotted on the left, followed by cp values of 0.001 and 0.005 respectively. Across the three plots, it is consistent

that the floor area has the greatest influence over the price of the resale flat, followed by: flat type, town, flat model and remaining lease. It can be seen that the variables ranking are the same for the optimal tree and cp value of 0.001. On the other hand, for the cp value of 0.005, the variable representing the distance of the nearest hawker centre now holds more importance than the variable reflecting the number of hospitals within a 2km radius.

Compared to the initial model, the floor area is shown to be more important than the flat type as it has yet to be discretized. After normalization, the variable importance ranking for remaining lease drops from 5th in the initial model to 6th or 7th in the improved models. This goes to show that discretization and normalization affects the evaluation and insights obtained from our model.

6.3 Random Forest Regression

Random tree regression was performed again with the new amenities data. The grid search parameters and hyperparameters were kept the same to allow for a fair comparison between the model with amenities data and the model without amenities data except for 'mtries' to reflect the increase in number of columns. The results on train data and test data are as seen in Figures 32 and 33 respectively and the respective R-squared values were 0.9491656 and 0.9581103.

```
H2ORegressionMetrics: drf
** Reported on training data. **
** Metrics reported on Out-Of-Bag training samples **

MSE:  992722668
RMSE:  31507.5
MAE:  22790.61
RMSLE:  0.0698859
Mean Residual Deviance :  992722668
```

Figure 32

```
H2ORegressionMetrics: drf

MSE:  809798323
RMSE:  28456.96
MAE:  20625.55
RMSLE:  0.0631138
Mean Residual Deviance :  809798323
```

Figure 33

6.3.1 Residual Analysis

Residuals are differences between the one-step predicted output from the model and the measured output from the validation data set. As shown in Figure 34 below, the residuals are quite symmetrically distributed, clustered around the lower single digits of the y-axis and there are no clear patterns in general. Hence, it can be concluded that the model is appropriate for the data.

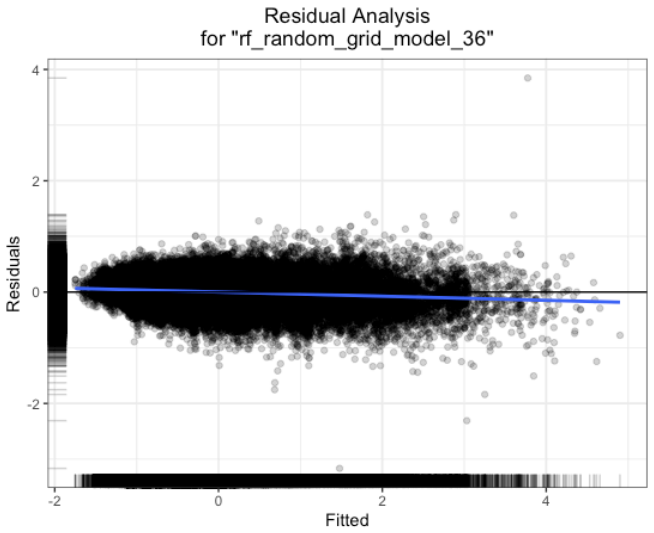


Figure 34

6.3.2 SHAP Summary Plot

Complex predictive models, such as random forest regression, are not easy to interpret. A recent technique to interpret black-box models has stood out among others: SHAP (SHapley Additive exPlanations) developed by Scott M.Lunderberg. Shapley value calculates the importance of a feature by comparing a model's prediction with and without the feature. However, since the order in which a model sees features can affect its predictions, this is done in every possible order so that the features are fairly compared. The variables are ranked in the descending order in the plot as seen in Figure 35. Most evidently, 'floor_area_sqm' had the largest effect, in both positive and negative directions, on the resale price, followed by flat_type and town. A simplified version is shown in Figure 36.

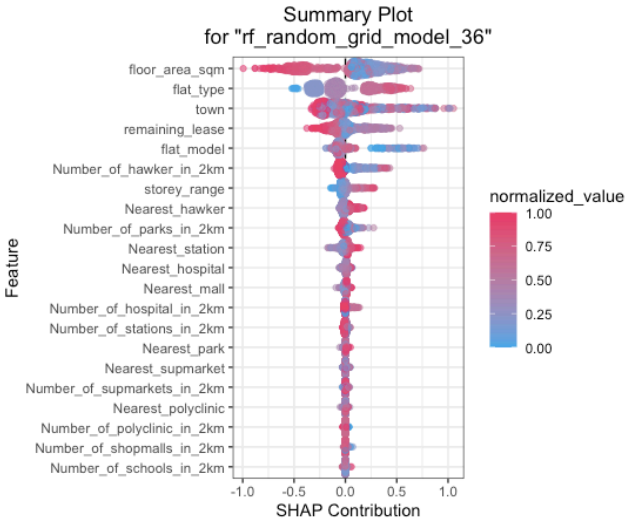


Figure 35

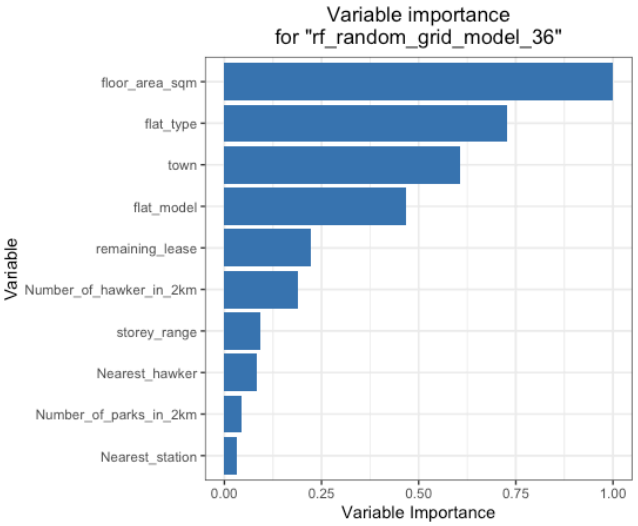


Figure 36

7. Analysis

7.1 Suggested Model

Amongst all the models generated, the random forest model is recommended to be used as the final model to predict HDB resale flat prices. In terms of error comparison, metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are used. MAE measures the absolute average distance between the actual and predicted data while RMSE takes the square root of the squared average distance between the actual and predicted data. Since RMSE is a quadratic score, it also penalises larger errors more. Both metrics measure the average prediction error ranging from 0 to infinity with negatively-orientated scores, which means the lower the error value the better the model.

Description	Linear Regression	CART Model	Random Forest
MAE	39,092.45	25,058.32	20,715.70
RMSE	51,631.82	34,303.19	28,443.35

Table 3

From Table 3 above, it is observable that Random Forest has the lowest MAE value of 20,715.70 and RMSE value of 28,443.35 (both rounded to 2d.p.) for the predicted resale price amongst the three models. As such, it can be concluded that it is the most accurate in terms of price predictions.

In addition, the strengths and weaknesses of the models generated are further compared and analysed. Linear regression is a strong model as it enables the interpretation of results by looking at the coefficients for each variable which encapsulates the magnitude of impact that each variable has. However, it assumes a linear relationship between the various variables and the related results, which may not be the case in reality. Moreover, certain results may be biased since the model is parametric in nature.

Conversely, non-parametric techniques such as CART do not need any assumptions for underlying distributions and are able to provide valuable insights into significant amounts of data as used in this project. However, it is fair to note that the tree becomes quite complex and interpretation of the results may not be intuitive.

In addition, a small change in the dataset can cause instability in the tree which can cause variance. This is where Random Forest comes in. Since it leverages the power of multiple decision trees, it does not just rely on the feature importance of just one decision tree and chooses features randomly during the training process. As such, data is better generalised, which achieves a higher accuracy. Furthermore, the large dataset means there are a diverse number of variables to consider, in which random forest would be a much better fit.

7.2 Addressing the Business Problem

To help the couples in planning their housing arrangement, an application fed with random forest models will be created. The application will be made very intuitive, where the couple can simply select the characteristics of the HDB resale flats that they are looking for, and the application will be able to generate a predicted HDB resale price based on their input.

Across the models generated, it is observable that 'variable floor area in square metres', 'flat type' and 'town' take precedence over the accessibility to each type of amenities such as 'food and retail' or 'healthcare' facilities.

As such, in terms of family planning purposes in the near future, the floor area and type of flat should be significantly considered. Buyers could consider going for a smaller floor area with a minimalist design and space saving furniture to reduce the amount they are paying for the resale flat. The type of flat type measures the number of rooms available and couples should conduct good family planning so that they only purchase the flats with the number of rooms they need.

For the variable 'town', couples have to weigh the pros and cons of living in certain areas as the location of the town significantly affects the resale price. Moreover, subsidies such as the Proximity Housing Grant (PHG), where couples live near their parents, are factors that also contribute to 'town's importance since it lifts a certain extent of couples' financial burdens. On the other hand, the latter variables take a relative backseat, so they may be considered less important. As such, the former variables should be the primary considerations for young couples and the remaining variables such as the relevant amenities may be considered secondary considerations.

Ultimately, the model provides many variables that potential buyers can consider and it is important to balance these variables with their financial budget for them to make an informed and practical decision.

7.3 Limitations to our Research & Analysis

In reality, there are many other factors which will affect the resale price, in which some of the information was difficult to obtain. For instance, the price of neighbouring flats in a time period, the geographic direction which the flat faces *(International Journal of Housing Market and Analysis)* and changes in government policies are just a number of examples that are difficult to quantify. Furthermore, the data only covered up to a certain period in time (2020), so there may be certain inaccuracies due to unforeseen events in the future. For instance, the data does not include the newly-open Thomson East-Coast MRT line. As such, there may be adjacent flats and towns that are now in more favourable positions, which may affect their resale flat prices.

8. Conclusion

Through the above analysis, we hope that young couples will be able to make a more informed decision regarding the future housing plans via the usage of the models and solutions that have been proposed. With a greater wealth of knowledge in their hands, young couples can better consider resale flats as an alternative to BTOs and in the event they decide to go for resale flats, to be able to better plan out their finances, and set realistic goals for their ideal HDB flat.

Despite the analysis in the report showing that the random forest model is the most suited for the created housing application, it is notable that there are still limitations involved, as well as a multitude of variables that are difficult to quantify. Hence, the proposed models should only serve as a guideline and should not be used as a replacement for thorough market research.

9. Appendices

Appendix A - Elaboration of the different flat models

Flat Model	Elaboration
2-room	Under HDB's the 2-room Flexi Flats (2-room) are 2 room flats mainly catered towards the elderly, who have the flexibility to choose the length of their lease on the flat based on eir age needs and preferences. For first- and second-timer families, they can buy these flats on a 99 year lease.
Adjoined Flat	Adjoined flats are two flats converted into a single property and usually consist of 4 or 5 bedrooms.
Apartment	Introduced in the 1990s, they were made to replace Model A with better quality furnishings.
DBSS	Under HDB's Design, Build and Sell Scheme (DBSS) which occurred from 2005 to 2012, DBSS flats are public apartments built by private developers, where the interior resembles a private condominium than a HDB flat
Improved	(For 1/2/3/4/5 room flat types) -- The Improved flat model was introduced in the 1960s, with 3 or 4 room flat types usually having a separate toilet and shower. While the naming convention has been dropped since the 2000s, it is still frequently displayed in resale flat prices.
Executive Maisonette	Before halting in the year 1995, Executive Maisonettes were constructed to be considered as the more premium version of Maisonettes.
Maisonette	Maisonette are two-storey HDB flats. However, they are no longer in production and are now mainly found in mature estates such as Ang Mo Kio, Bedok, Choa Chu Kang, Hougang, Pasir Ris, Queenstown and Sembawang.
Model A	Introduced in the 1970s, Model A usually had two inbuilt full-sized toilets. While the naming convention has been dropped since the 2000s, it is still frequently displayed in resale flat prices.
Model A2	Similar to Model A, but Model A2 was only found in Selective En bloc Redevelopment Scheme (SERS), which is a scheme
Multi Generation	Multi Generation flats can only be bought by multi-generational families, consisting of a married couple and one set of parents.

New Generation	Introduced between the 1970s and 80s, New Generation flats were usually 3 or 4 room flats with two toilets
Premium Apartment	A HDB flat which comes with an additional space.
Simplified	Introduced between the 1970s and 1980s, Simplified flats were usually 3 or 4 room flats with two toilets
Standard	(For 2/3/4/5 room flat types) Introduced between the 1960s and 1970s, Standard flats usually had a toilet and shower in the same room for 2 to 4 room flats, and had two toilets for 5 room flats
Terrace	Terrace HDB houses come in two storeys and a private garden and were first constructed in the 1950s. Most of them are considered to be categorised under 3 room flats, despite have a significantly larger floor area
Type S1S2	In the 2000s, HDB introduced the sale of apartment units under its sales exercise for ThePinnacle@Duxton, which was HDB's first 50 storey integrated housing development. For application of HDB policies, S1 and S2 apartments were treated as 4-room and 5-room flats respectively.

Appendix B: Visualisations

Figure 1

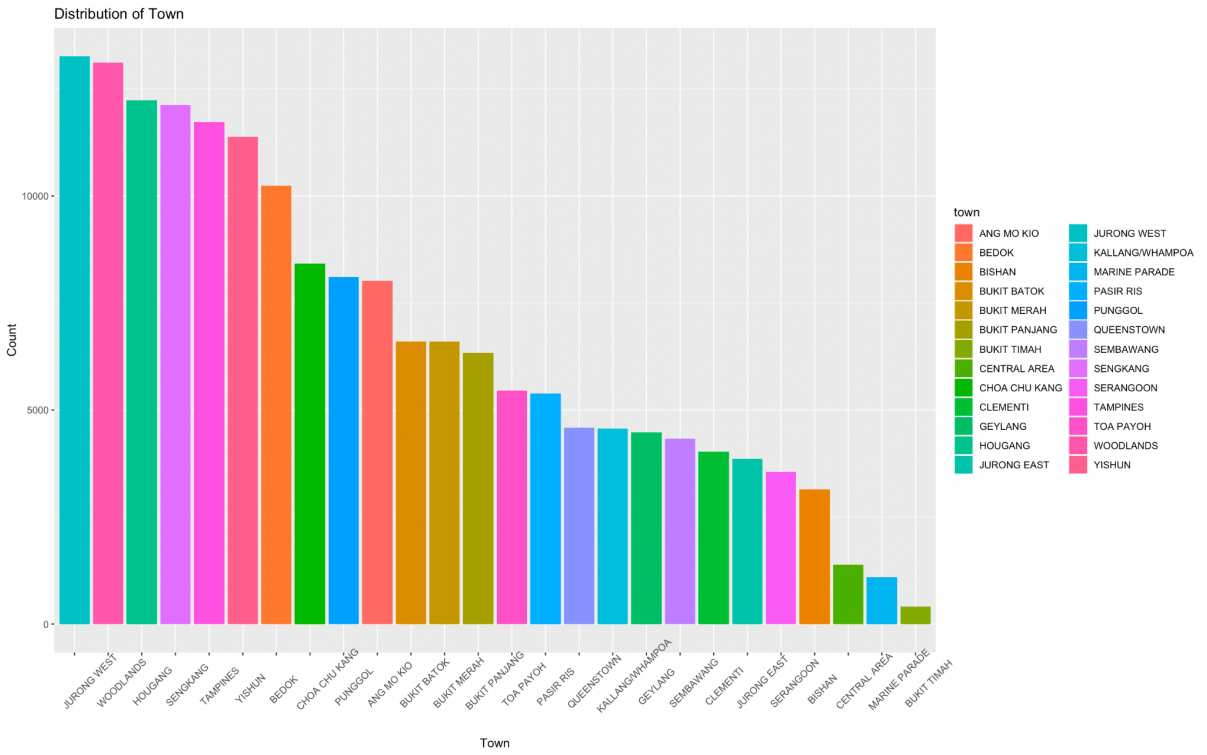


Figure 2

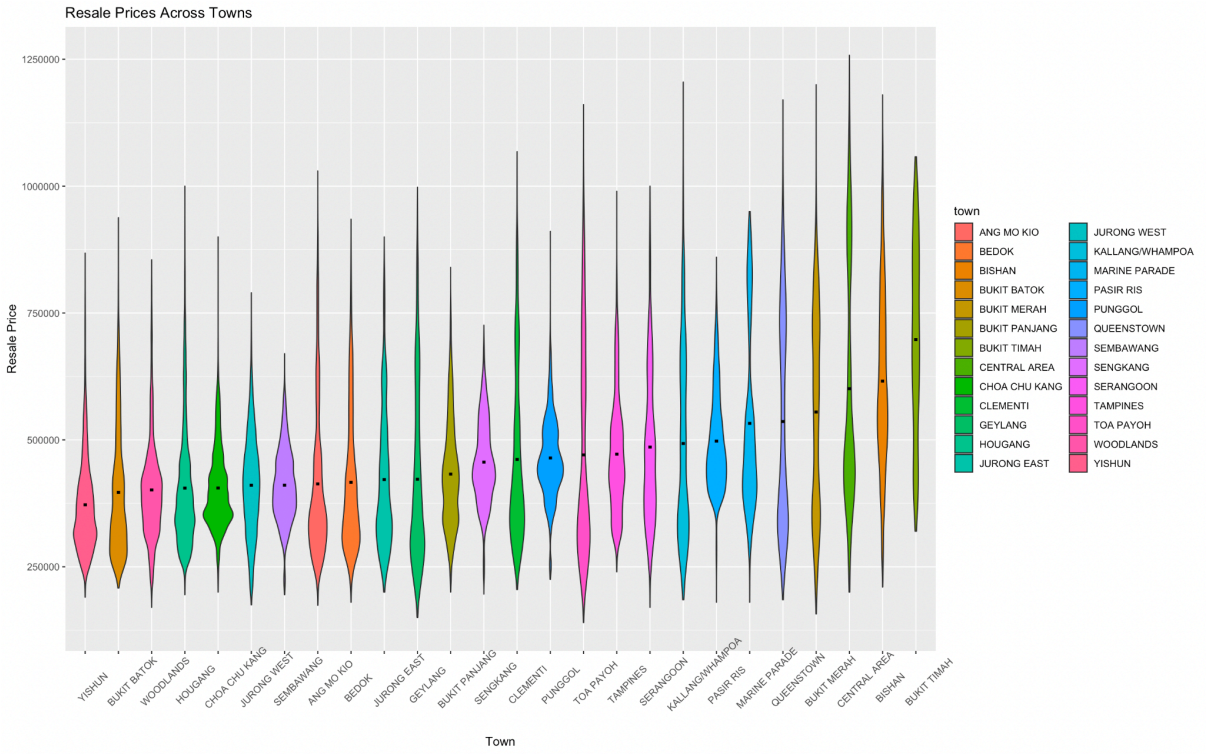


Figure 3

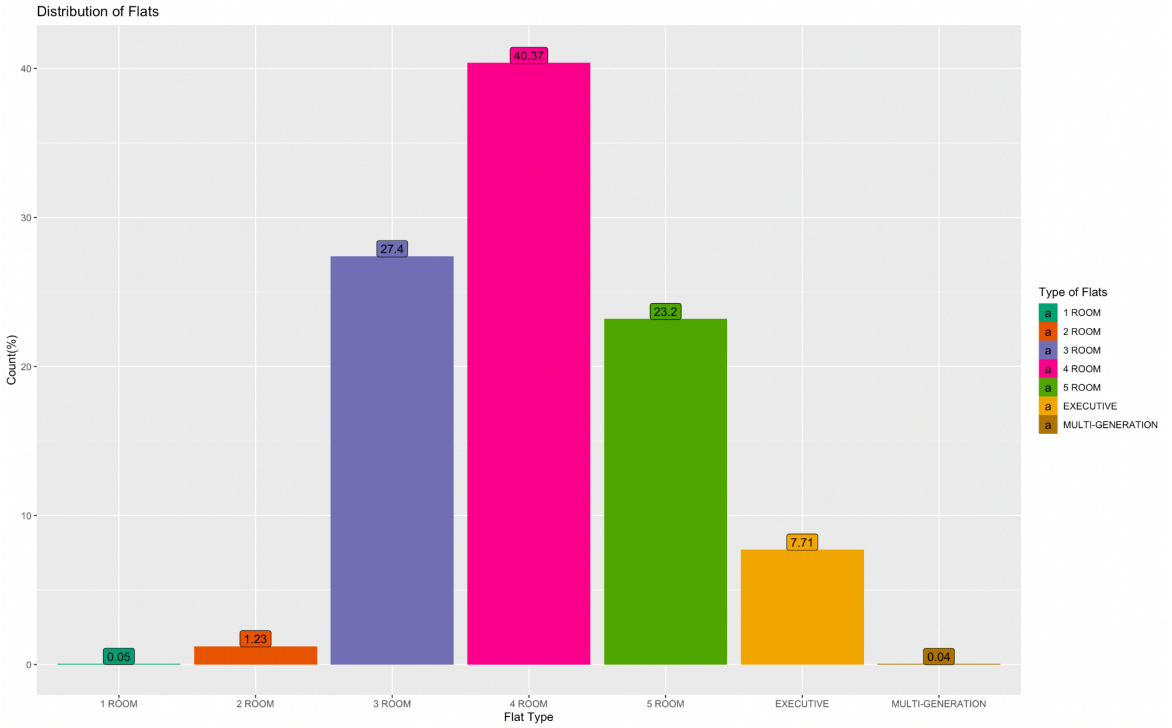

Distribution of Flats
Count(%)
40
30
20
10
0
40.37
27.4
23.2
7.71
1.23
0.05
0.04
1 ROOM
2 ROOM
3 ROOM
4 ROOM
5 ROOM
EXECUTIVE
MULTI-GENERATION
Flat Type
Type of Flats
a 1 ROOM
a 2 ROOM
a 3 ROOM
a 4 ROOM
a 5 ROOM
a EXECUTIVE
a MULTI-GENERATION

Figure 4

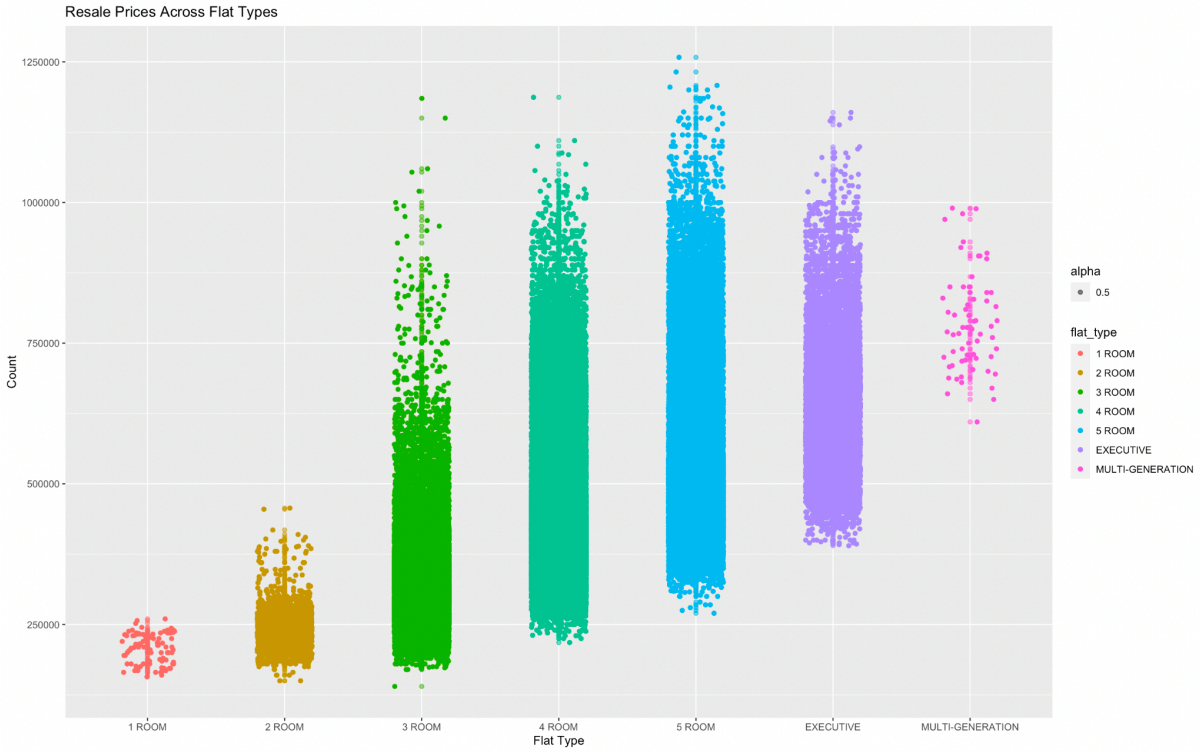


Figure 5

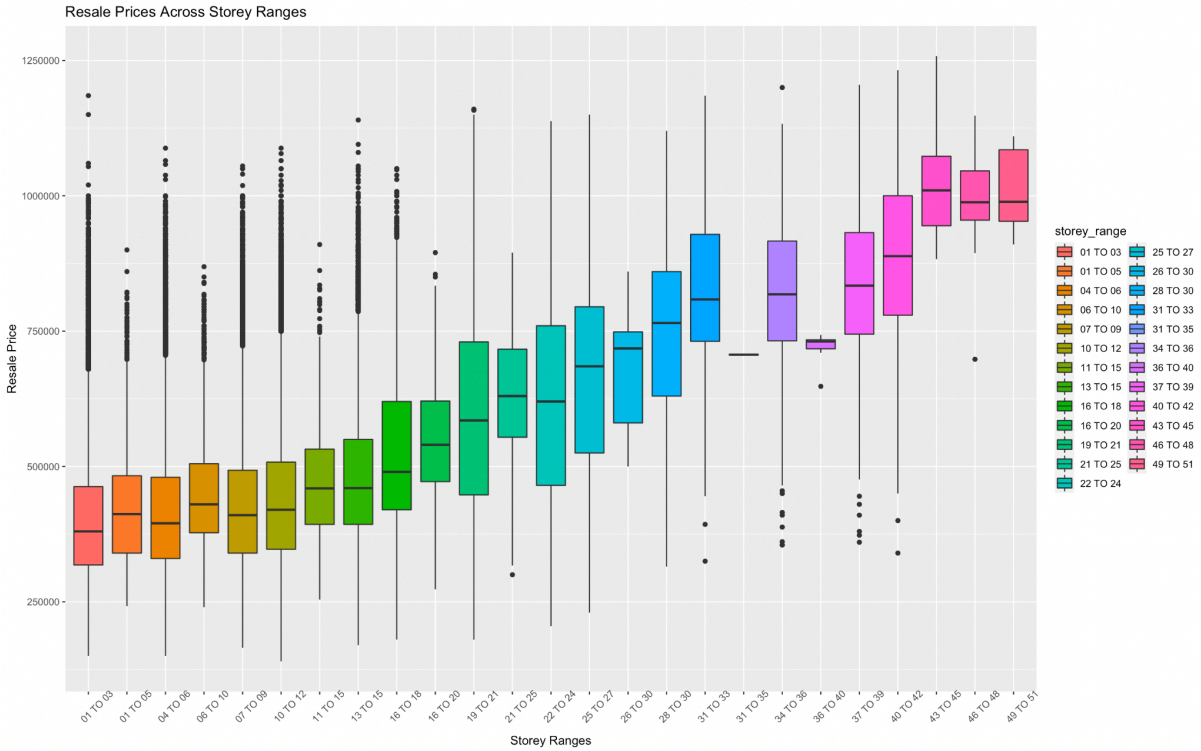


Figure 6

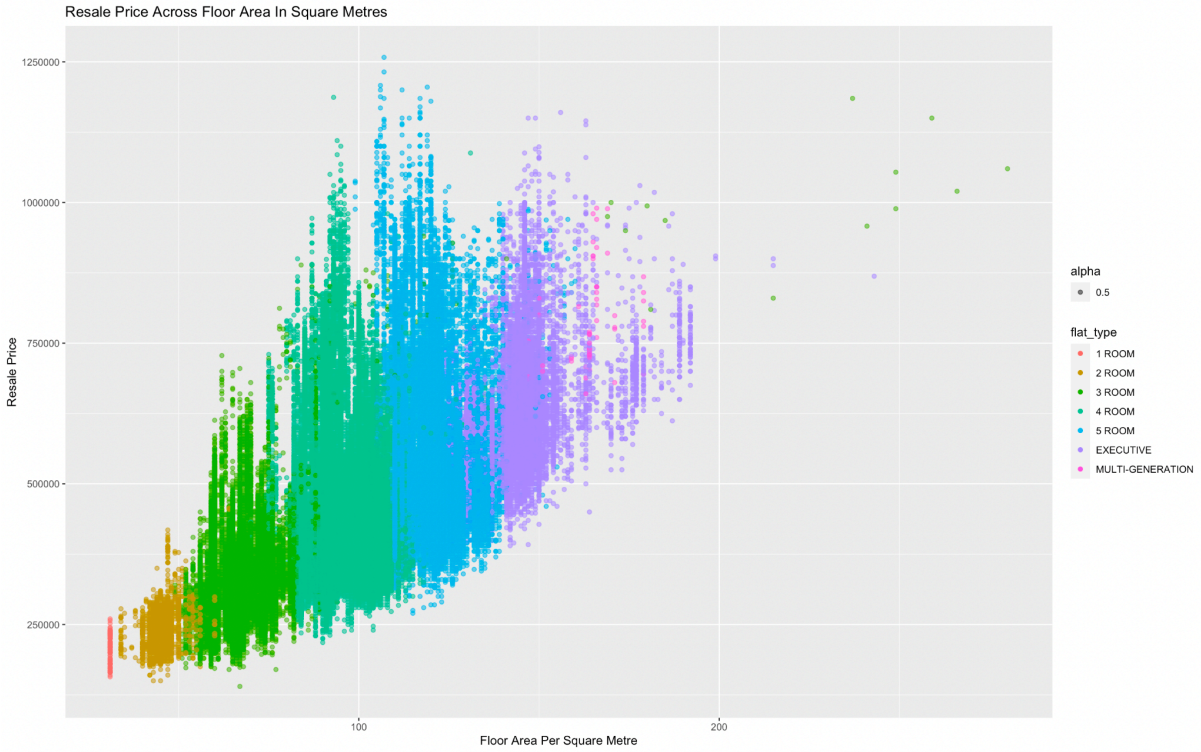


Figure 7

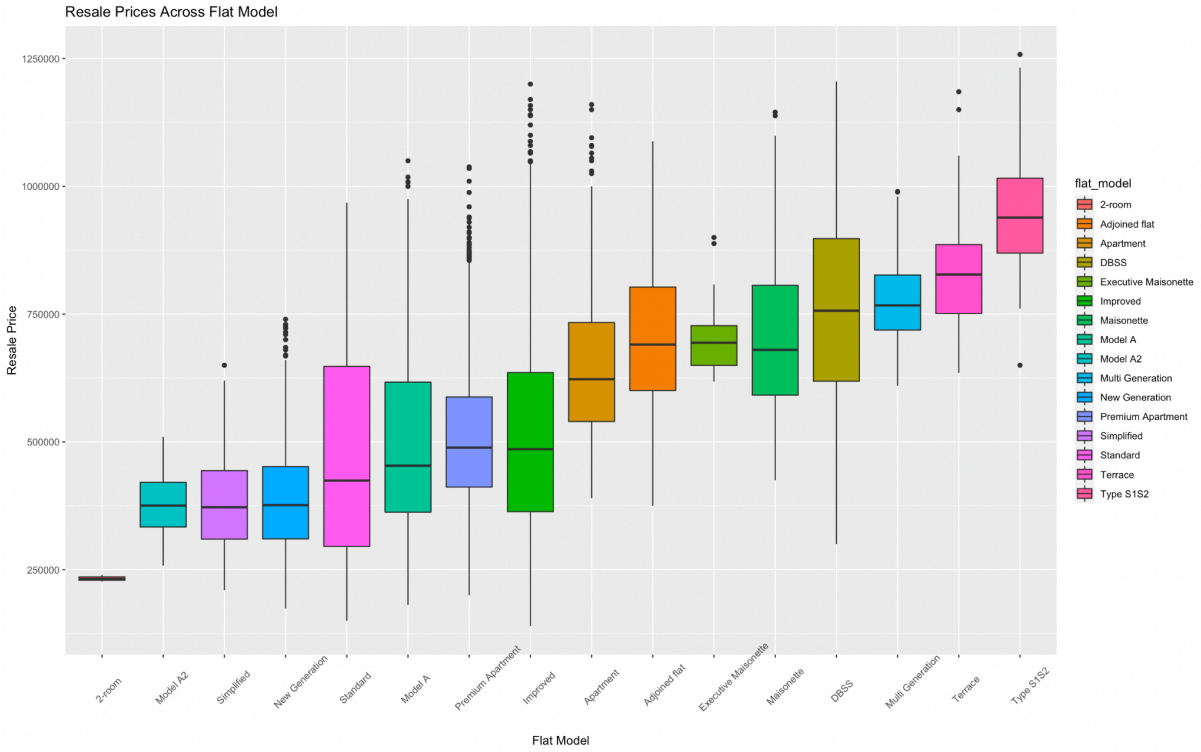


Figure 11

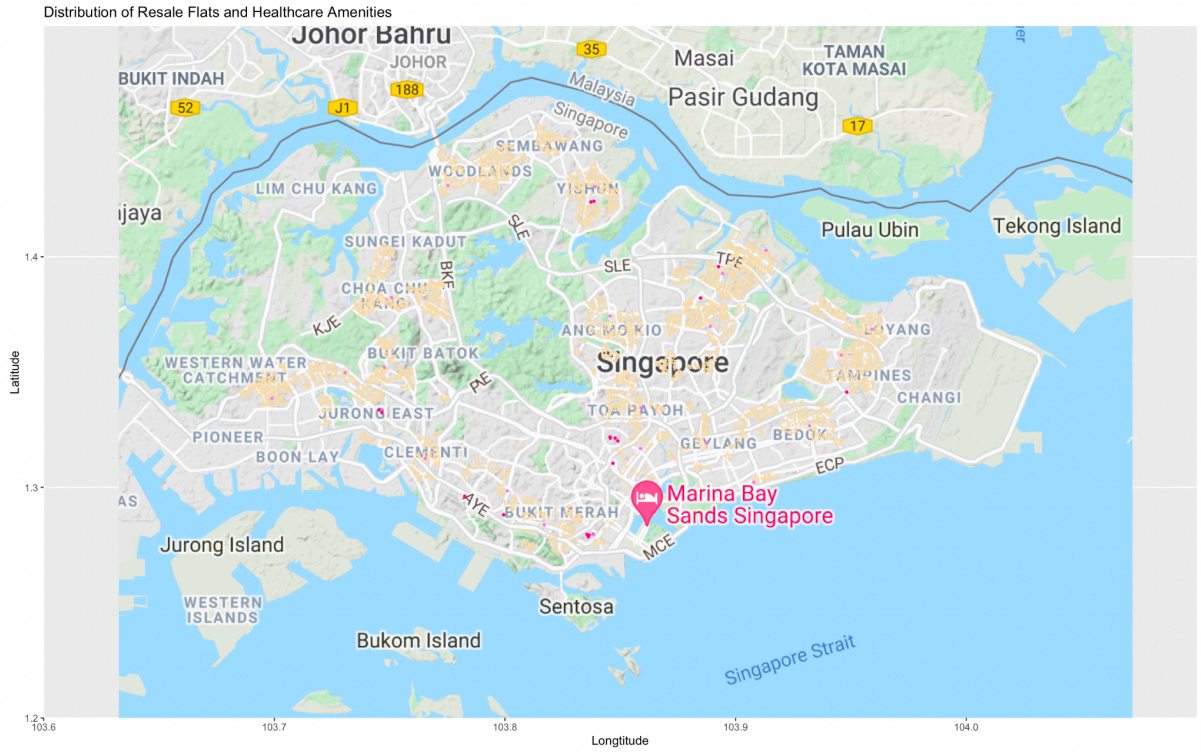


Figure 12

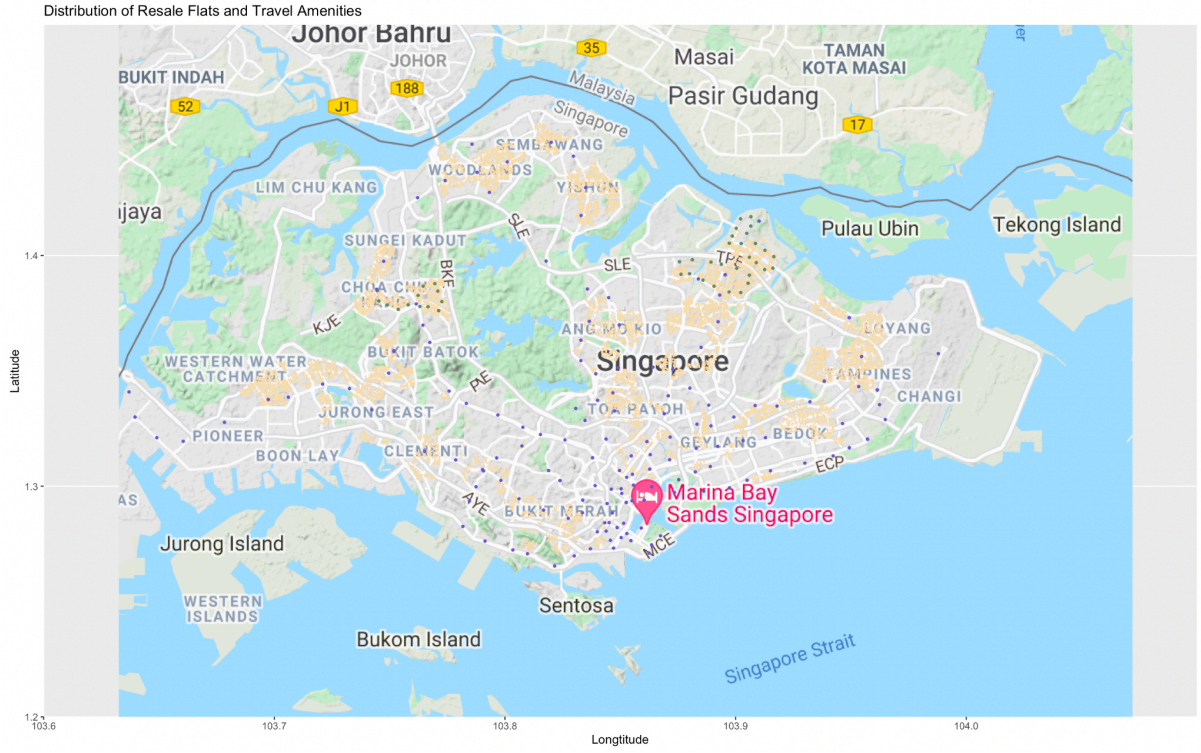


Figure 13

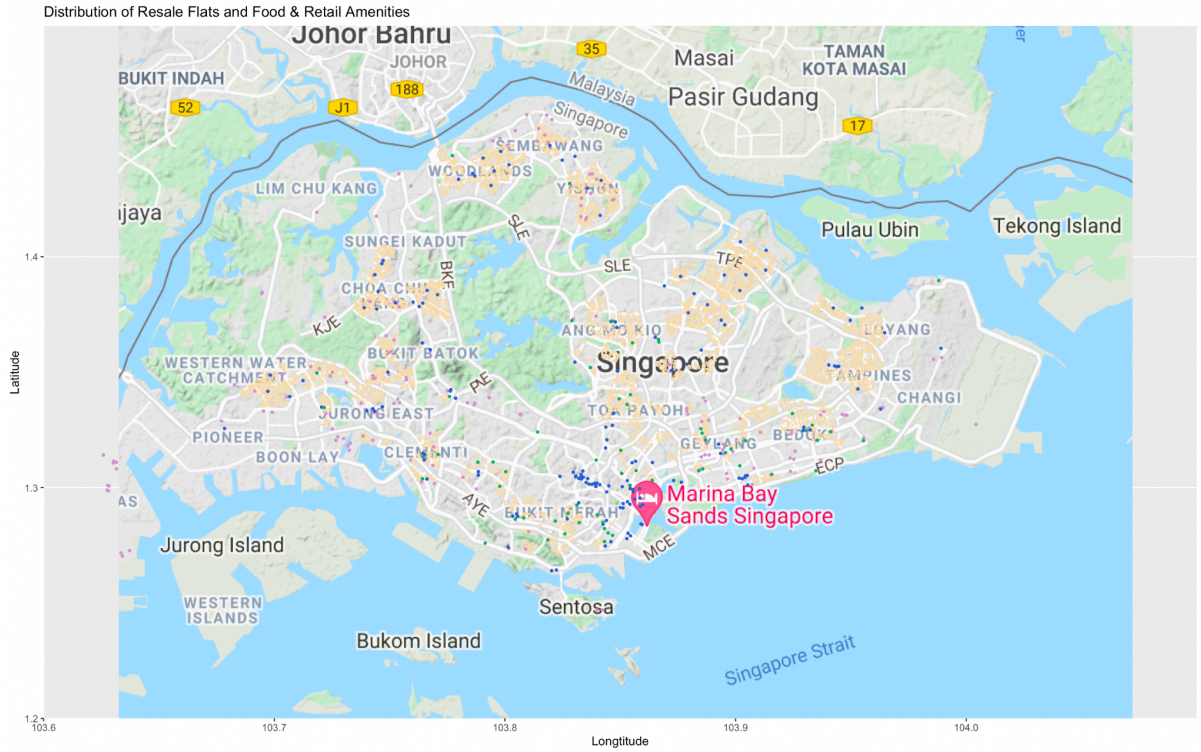


Figure 14

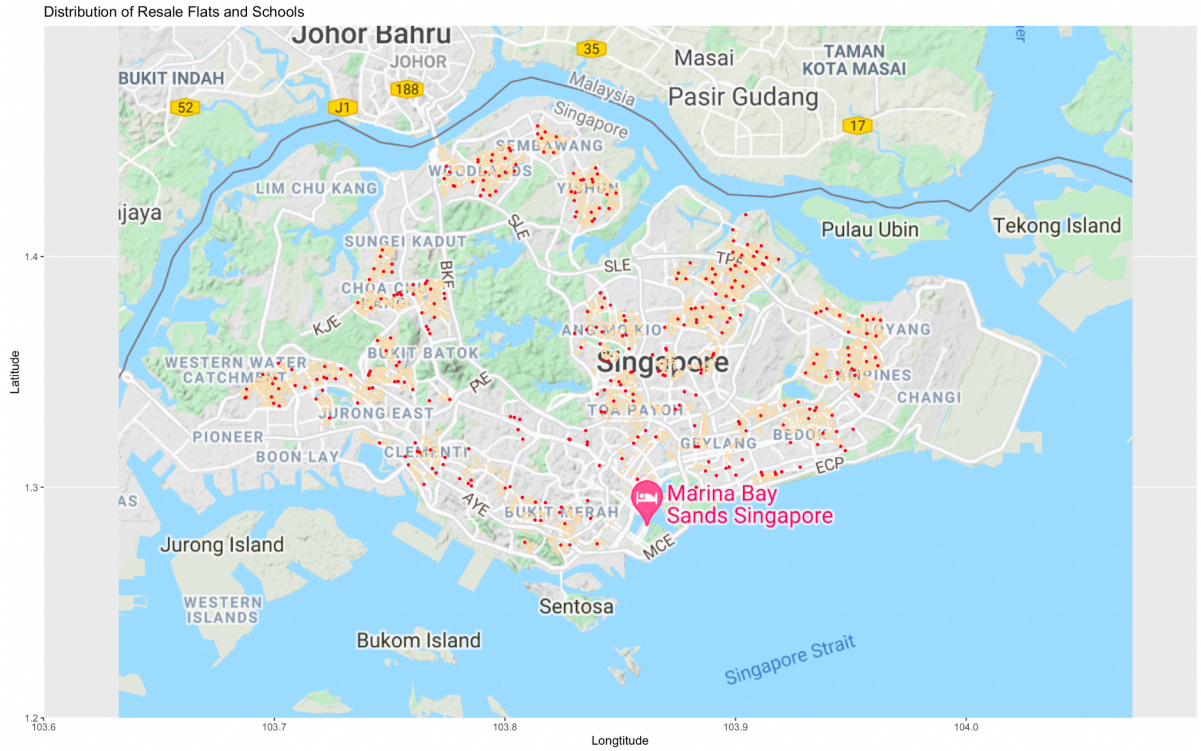
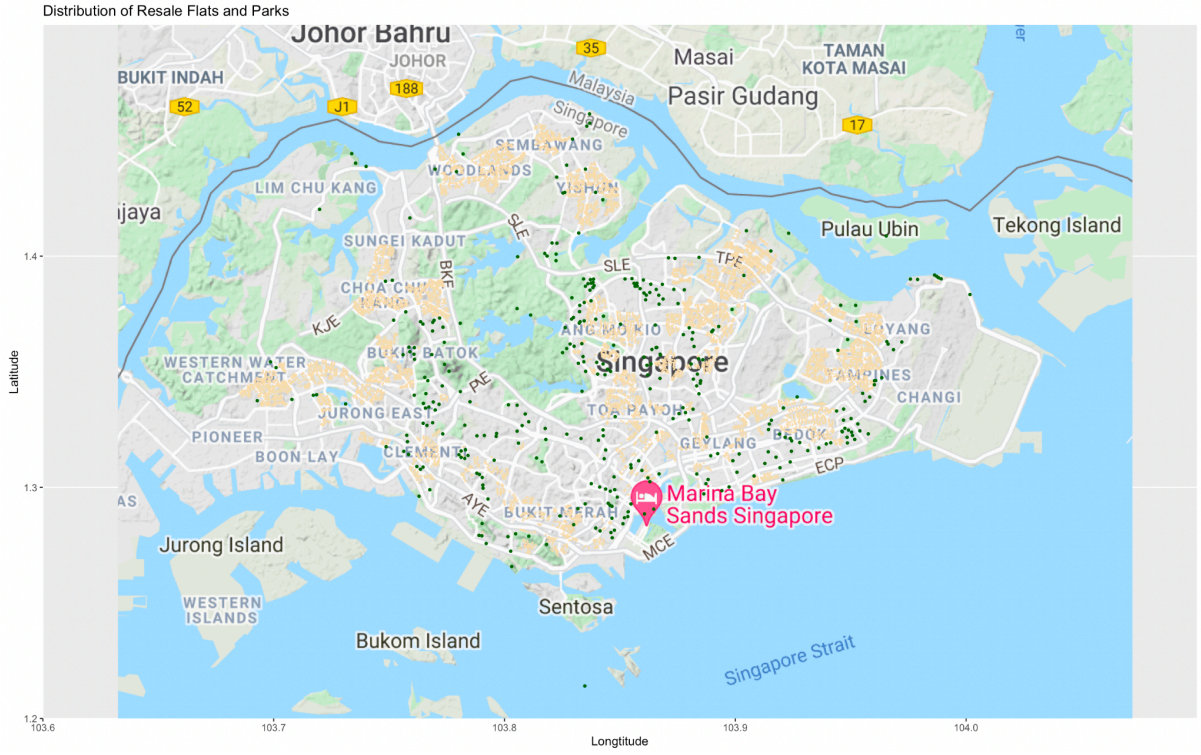


Figure 15


Distribution of Resale Flats and Parks
Johor Bahru
JOHOR
BUKIT INDAH
52
J1
188
35
Masai
TAMAN KOTA MASAI
Pasir Gudang
17
Malaysia
Singapore
SEMBAWANG
WOODLANDS
YISHUN
LIM CHU KANG
SUNGEI KADUT
SLE
BKE
KJE
CHOA CHU KANG
ANG MO KIO
TPE
Pulau Ubin
Tekong Island
WESTERN WATER CATCHMENT
BUKIT BATOK
Singapore
TAMPINES
CHANGI
PIE
JURONG EAST
TOA PAYOH
PIONEER
BOON LAY
CLEMENTI
GEYLANG
BEDOK
ECP
AYE
BUKIT MERAH
Marina Bay Sands Singapore
MCE
Jurong Island
WESTERN ISLANDS
Sentosa
Bukom Island
Singapore Strait
Latitude
1.4
1.3
1.2
103.6
103.7
103.8
103.9
104.0
Longtitude

10. References

1. Housing & Development Board, (H. D. B.) (2021). Public housing – a Singapore icon. HDB. Retrieved November 3, 2021, from https://www.hdb.gov.sg/about-us/our-role/public-housing-a-singapore-icon.
2. Property Guru. (2017). What is a BTO flat? | propertyguru Singapore. What is a BTO flat? Retrieved November 3, 2021, from https://www.propertyguru.com.sg/property-guides/what-is-a-bto-flat-2-6515.
3. Mae, C. S. (2021, August 12). [2021 BTO application edition] step-by-step guide to buying a HDB BTO flat. DollarsAndSense.sg. Retrieved November 3, 2021, from https://dollarsandsense.sg/bto-application-edition-step-step-guide-buying-hdb-bto-flat/.
4. PropertyGuru. (2021, February 1). The ultimate guide to buying a resale HDB flat in 2017. Property Blog Singapore - Stacked Homes. Retrieved November 3, 2021, from https://test-apr21.stackedhomes.com/editorial/the-ultimate-guide-to-buying-a-hdb-resale-flat-in-2017/.
5. Housing & Development Board (HDB). (2021). Flat Supply & Applications received. HDB. Retrieved November 3, 2021, from https://services2.hdb.gov.sg/webapp/BP13BTOENQWeb/AR_Aug2021_BTO?strSystem=BTO.
6. HomeRenoGuru. (2021, February). BTO vs resale: 9 factors to consider [+sale of balance option]. HomeRenoGuru. Retrieved November 3, 2021, from https://www.homerenoguru.sg/articles/tips-advice/bto-vs-resale/.
7. UCLA Institute for Digital Research & Education. (2021). Factor Variables. IDRE Stats. Retrieved November 3, 2021, from https://stats.idre.ucla.edu/r/modules/factor-variables/.
8. Flynn, T. D. (2014, August 1). Why we should account for inflation. Harvard Business Review. Retrieved November 3, 2021, from https://hbr.org/1977/09/why-we-should-account-for-inflation.
9. Pinky. (2020, August 5). Mature vs non-mature estates in Singapore: Which is better (for you)? Carousell Property. Retrieved November 3, 2021, from https://blog.carousell.com/property/mature-non-mature-estates-hdb-singapore/.
10. Visit Singapore. (2021). Orchard Road. A shopping paradise - Visit Singapore Official Site. Retrieved November 3, 2021, from https://www.visitsingapore.com/see-do-singapore/places-to-see/orchard/.
11. The Business Times (2021, September 21) *Is the double-digit growth iof HDB resale flat prices sustainable?* Businesstimes.com.sg. Retrieved November 3, 2021, from https://www.businesstimes.com.sg/hub-projects/property-2021-sept-issue/is-the-double-digit-growth-of-hdb-resale-flat-prices.
12. The Straits Times (2021, August 17). *Hougang BTO flats draw more than 10,000 applicants, all seven projects oversubscribed*. The Straits Times. Retrieved November 3, 2021, from

https://www.straitstimes.com/singapore/housing/hougang-bto-flats-draw-more-than-10000-applicants-all-seven-projects.
13. CNA (2021, May 3). The big read: Rising prices, building delays - young couples face Perfect storm in quest for home sweet home. CNA. Retrieved November 3, 2021, from https://www.channelnewsasia.com/singapore/big-read-rising-property-prices-building-delays-young-couples-1338166.
14. Singh, M. (2021, January 11). *Advantages and disadvantages of linear regression, its assumptions, evaluation and implementation*. Medium. Retrieved November 4, 2021, from https://manish-ks.medium.com/advantages-and-disadvantages-of-linear-regression-its-assumptions-evaluation-and-implementation-61437fc551ad.
15. Ojha, A. K. (2017, May 16). *Use a classification and regression tree (CART) for Quick Data insights*. iSixSigma. Retrieved November 4, 2021, from https://www.isixsigma.com/methodology/lean-methodology/use-a-classification-and-regression-tree-cart-for-quick-data-insights/.
16. Yadav, A. (2019, January 11). *Decision trees*. Medium. Retrieved November 4, 2021, from https://towardsdatascience.com/decision-trees-d07e0f420175.
17. Housing, D. B. (2021). *Living with/ near parents or child*. HDB. Retrieved November 4, 2021, from https://www.hdb.gov.sg/residential/buying-a-flat/resale/financing/cpf-housing-grants/living-with-near-parents-or-child.
18. Housing Development Board, (H. D. B. (2021). *2-room flexi flats*. HDB. Retrieved November 4, 2021, from https://www.hdb.gov.sg/residential/buying-a-flat/new/eligibility/2room-flexi-flat.
19. Tan, R. (2018). *The different types of HDB houses you can call home*. Yahoo! News. Retrieved November 4, 2021, from https://sg.news.yahoo.com/different-types-hdb-houses-call-020000642.html.
20. Property, G. (2020). *What is a DBSS flat and is it worth buying? | propertyguru ...* PropertyGuru. Retrieved November 4, 2021, from https://www.propertyguru.com.sg/property-guides/dbss-singapore-17893.
21. Alida, T. (2021, September 20). *HDB flat types, 3STD, 3NG, 4S, 4A, 5i, EA, Em, Mg, etc*. The world of Teoalida. Retrieved November 4, 2021, from https://www.teoalida.com/singapore/hdbflattypes/.
22. H, D. B. (2021). *SERS*. HDB. Retrieved November 4, 2021, from https://www.hdb.gov.sg/residential/living-in-an-hdb-flat/sers-and-upgrading-programmes/sers.
23. Guru, P. (2021). *HDB terrace houses: 6 of these public ... - propertyguru*. Retrieved November 4, 2021, from https://www.propertyguru.com.sg/property-guides/cheap-hdb-terrace-house-under-850k-35237.
24. J, R. (2021, June 28). *HDB landed terrace houses: 5 factors you must consider before buying one*. Property Blog Singapore - Stacked Homes. Retrieved November 4, 2021, from

https://stackedhomes.com/editorial/hdb-landed-terrace-houses-5-factors-you-must-consider-before-buying-one/.
25. *HDB launches 50-storey The Pinnacle @ Duxton today*. Getforme Singapore HDB launches 50-storey the pinnacle @ duxton frontpage edition 29 May 2004. (2004). Retrieved November 4, 2021, from http://getforme.com/previous2004/previous290504_hdblaunchesthepinnancleatduxton.htm.
26. HDB. (2021). *Living with/ near parents or child*. HDB. Retrieved November 4, 2021, from https://www.hdb.gov.sg/residential/buying-a-flat/resale/financing/cpf-housing-grants/living-with-near-parents-or-child.
27. Casas, P. (2019, March 19). *How to interpret shap values in R (with code example!)*. Data Science Heroes Blog. Retrieved November 4, 2021, from https://blog.datascienceheroes.com/how-to-interpret-shap-values-in-r/.
28. *Qualtrics*. Qualtrics XM. (2021, April 12). Retrieved November 4, 2021, from https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/.
29. HDB. (2021). *An HDB flat for your different life cycle needs*. HDB. Retrieved November 4, 2021, from https://www.hdb.gov.sg/about-us/news-and-publications/publications/hdbspeaks/an-hdb-flat-for-your-different-life-cycle-needs#:~:text=What%20is%20the%20Length%20of,recycled%20to%20house%20future%20generations.
30. -, P. (2020, August 5). *Mature vs non-mature estates in Singapore: Which is better (for you)?* Carousell Property. Retrieved November 4, 2021, from https://blog.carousell.com/property/mature-non-mature-estates-hdb-singapore/.
31. Prabhakaran, S. (2021, June 5). *Linear regression - a complete introduction in R with examples*. Machine Learning Plus. Retrieved November 4, 2021, from https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/.
32. *Explanation of Decision Tree Model*. Explanation of the decision tree model. (2021). Retrieved November 4, 2021, from https://webfocusinfocenter.informationbuilders.com/wfappent/TLs/TL_rstat/source/DecisionTree47.htm.
33. Frost, J., Rodrigues, P., Singh, J., Hardy, S., & Nousheen, R. (2021, September 24). *Multicollinearity in regression analysis: Problems, detection, and solutions*. Statistics By Jim. Retrieved November 4, 2021, from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/.
34. bhandari, aniruddha. (2020, April 16). *Multicollinearity: Detecting multicollinearity with VIF*. Analytics Vidhya. Retrieved November 4, 2021, from https://www.analyticsvidhya.com/blog/2020/03/what-is-multicollinearity/.
35. Vashisht, P. author B. R. (2021, September 20). *Machine learning: When to perform a feature scaling?* atoti. Retrieved November 4, 2021, from https://www.atoti.io/when-to-perform-a-feature-scaling/.

36. Bronshtein, A. (2020, March 24). *Train/test split and cross validation in Python*. Medium. Retrieved November 4, 2021, from https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6.
37. Cahya, kamenrider. (2018, March 11). *Cross-validation essentials in R*. STHDA. Retrieved November 4, 2021, from http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/#discussion.
38. Prabhakaran, S. (2021, June 5). *Linear regression - a complete introduction in R with examples*. Machine Learning Plus. Retrieved November 1, 2021, from https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/.
39. March 2008 International Journal of Housing Markets and Analysis 1(1):81-101